Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

绪论：

训练深度神经网络由于每层输入的分布在训练期间改变的事实而变得复杂，例如先前层参数的改变。这需要低的学习率和仔细的参数初始化来减慢训练速度，众所周知训练饱和非线性模型是比较困难的。我们把这种现象称为内部协变量转移，并通过归一化每层的输入来解决这个问题。我们的方法通过归一化模型结构的一部分并且对每个训练小批量执行归一化来实现其强度。批量归一化允许我们使用很高的学习率并且对初始化不太在意。它也作为一个正则化，在一些情况下消除了对Dropout的需要。应用于最先进的图像分类模型，批量归一化在减少了14倍训练步骤的情况下实现了相同的精度，并以显著边缘击败了原始模型。使用批量归一化网络的一个集合，我们提高了在ImageNet分类上发布的最佳结果:获得了4.9%前5验证误差（和4.8%测试误差），超出了人类评估者的准确率。

简介：

深度学习极大地提升了视觉，语言和许多其他领域的最先进技术。随机梯度下降已经被证明是训练神经网络一个有效的方法，并且随机梯度下降变量如动量和Adagrad已经被用来去获得最先进的性能。随机梯度下降优化网络的参数，以便最小化损失



其中是训练集。训练按步骤进行，在每一步我们考虑一个大小为的小批量。这个小批量被用来去近似相关参数的损失函数的梯度，通过计算



使用小批量的样本，而不是一次一个样本，在几个方面有有帮助的。首先，小批量上的损失梯度是训练集上梯度的一个估计，其质量随着批量大小的增加而提高。第二，由于现代计算平台提供的并行性，对一个批量的计算比各个样本的个计算更有效。

当随机梯度是简单有效的，它需要仔细调整模型超参数，特别是使用在优化中的学习率以及模型参数的初始值。由于每层的输入受所有先前层的参数影响的事实，使训练复杂化，以致于网络参数的小变化随着网络变得更深而放大。

由于层需要连续的适应新的分布，层输入的分布的变化提出了一个问题。当输入分布到学习系统改变时，也就是说经历协变量移位，这个通常通过领域适应来处理。但是，协变量移位的概念可以作为一个整体延伸超出学习系统，适用于他自身的部分，比如子网络或者一个层。考虑一个网络的计算



其中和可以任意变换，参数学习去最小化损失。学习可以被视为如果输入如果输入被送入到子网络中



例如，一个梯度下降步骤



（批量大小为m学习率）完全等同于一个输入为的独立网络。因此，输入分布属性使得训练更有效—比如在训练和测试数据之间有相同的分布—也适用于子网络的训练。因此有利于的分布随时间保持不变。然后不必调整以补偿分布的变化。

固定一个子网络输入的分布将对子网络外的层产生积极的影响。用一个sigmoid激活函数考虑一个层，其中是层输入，权重矩阵和阈值向量是学习的层参数， 。随着增加，趋向于0.这意味着对于的所有维度，除了那些具有小的绝对值的，梯度流向下到会消失并且模型训练会变慢。但是因为是受，和下面所有层参数的影响，在训练期间对这些参数的改变将可能将的许多维度移动到非线性的饱和状态并且收敛减慢。这种效果是随着网络的深度的增加而放大。在实际应用中，饱和问题和导致的消失梯度通常通过使用Rectified Linear Units（ReLU）来解决。，仔细的初始化和小的学习率。然而，如果我们可以确保非线性输入的分布在网络训练时保持更加稳定，那么优化将不太可能在饱和状态中停滞，并且训练将加速。